

ABSTRACT

A system and method is provided for handling the collation of linguistic symbols of different languages that may have various types of compressions (e.g., from 2-to-1 to 8-to-1). A symbol table of the symbols identified as Unicode code points is generated, with each symbol tagged with a highest compression type of that symbol by sorting the compression tables of the various languages. During a sorting operation with respect to a given string, the tag of a symbol in the string is checked to identify the highest compression type of compressions beginning with that symbol, and the compression tables for the language with compression types equal or lower than the highest compression type of the symbol are searched using a binary search method to find a matching compression for the symbols in the string. A common search module is used to perform binary searches through compression tables of different compression types.